bio-protocol Preprint

Not Peer Reviewed
Ask a question online

An optimized four-step genome assembly strategy

Shilin Tian Huabin Zhao

Updated date: May 10, 2023

An abbreviated version of this protocol was published in Science Advances in May 2023

Comparative analyses of bat genomes identify distinct evolution of immunity in Old World fruit bats

DOI: 10.1126/sciadv.add0141

Detailed protocol

The protocol of our new assembly pipeline aims to ensure the generation of the extremely complete genome by taking advantage of this feature of Hi-C reads. The pipeline consists of four main steps (**Figure 1**):

1. Assemble and correct contigs

We used the high-quality Nanopore sequences to assemble initial contigs of *C. sphinx* by applying a 'correct-then-assemble' strategy from the package NextDenovo v2.4.0 with the parameters of 'read_cutoff = 1k, seed_cutoff = 32k, blocksize = 3g'. Subsequently, the initial contigs were corrected using the package NextPolish v1.3.1 with the Illumina paired-end reads and Nanopore sequences with the 'best' algorithm module. On the basis of the above analytical strategies, we obtained all contigs, namely, the Contig v1 assembly genome.

2. Cluster contigs by Hi-C interaction pair

The Hi-C read pairs were mapped to Contig v1 using the Bowtie2 software with a single-ended model. We obtained the unique mapped pairs and then discarded the invalid self-ligated and unligated fragments using the HiCUP pipeline (version 0.8.0). Finally, we used the valid interaction pairs to calculate the linkage frequency among all contigs using an agglomerative hierarchical clustering algorithm. We clustered the linked contigs based on the linkage suggested by Hi-C signal density, presumably along the same chromosome according to a preset number of partitions.

3. Classify the Nanopore reads to perform local assembly

We realigned the Nanopore reads to Contig v1 using the Minimap2 package. After removing the suboptimal alignment reads, we extracted the mapped reads of each linked contig group. Subsequently, we performed local assembly for each classified mapped read. Compared to global assembly, the strategy could avoid the false overlap relationships induced by repetitive sequences of the genome when constructing the string graph during assembly. We set an appropriate parameter 'seed_cutoff' for each local assembly according to the calculated results by the command 'seq_stat'. Then, the assembly and correction method in this step was similar to that mentioned in the step 1. Finally, we obtained a second set of all contigs, namely, Contig v2.

4. Anchor contigs onto chromosomes

Chromosome-scale genome was anchored by linkage information, restriction enzyme site, and string graph formulation using the algorithm ALLHiC. The placement and orientation errors exhibiting obvious discrete chromatin interaction patterns were manually adjusted.

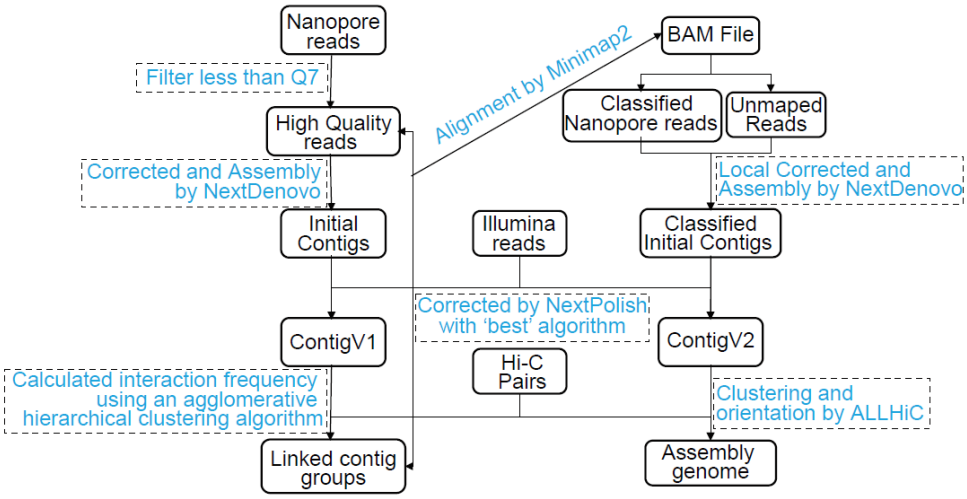

Figure 1. Schematic depiction of the genome assembly pipeline.

How to cite:(Readers should cite both the Bio-protocol preprint and the original research article where this protocol was used)

1. Tian, S. and Zhao, H. (2023). An optimized four-step genome assembly strategy. Bio-protocol Preprint. bio-protocol.org/prep2286.
2. Tian, S., Zeng, J., Jiao, H., Zhang, D., Zhang, L., Lei, C., Rossiter, S. J. and Zhao, H.(2023). Comparative analyses of bat genomes identify distinct evolution of immunity in Old World fruit bats. Science Advances 9(18). DOI: 10.1126/sciadv.add0141